GeoStream: Spatial Information Indexing Within Textual Documents Supported by a Dynamically Parameterized Web Service

Sallaberry, C., Royer, A., Loustau, P., Gaio, M., Joliveau, T., Le Ny, P-A.

Abstract Cultural heritage content is everywhere on the web: digital libraries, archives, and portals of museums or galleries. Cultural heritage document collections are characterized by contents related to a territory and its land's history. In this context, the GeoTopia project is supported by the CNRS-TGE-Adonis and focuses on archive data sharing and interpretation. It consists in a Content Management System (CMS) that aims to manage a repository of multimedia digital documents: it exploits information like origin, theme, period, area, etc. to index and/or query documents.
Our contribution is dedicated to spatial information contained in non structured textual documents. More specifically, we have developed a process flow that can extract the spatial information contained in textual documents. This process flow indexes spatial information and computes precise geolocalized representations. We propose to encapsulate it into the GeoStream specific web service and to make its behavior dynamically customizable for easier integration into such platforms used for the management of cultural heritage electronic documents.

1 Introduction

Managing electronic versions of archive data (histories, travelogues, stamps, maps, etc.) is becoming a task of major importance. In text oriented communities as li-

Sallaberry, C., Royer, A., Loustau, P., Gaio, M.,
LIUPPA, Université de Pau et des Pays de l'Adour, Avenue de l'université, BP 1155, 64013 PAU CEDEX, France, e-mail: christian.sallaberry@univ-pau.fr
and Joliveau, T.
CRENAM-ISIG CNRS/UMR 5600, Université Jean Monnet - Saint-Etienne, 42023 Saint-Etienne Cedex 02, France, e-mail: thierry.joliveau@univ-st-etienne.fr
and Le Ny, P-A,
Makina Corpus, 30 rue des Jeûneurs, 75002 Paris, France, e-mail: pierre-andre.le-ny@makina-corpus.com

braries, archives, museums etc. the geographic association of information (location, address, place name etc.) appears to be as important as temporal characterization [1]. At the same time Geomatics techniques are not any longer limited to topographical entities, material objects or biophysical processes. They are involved in building more general Georeferenced Information Systems including all kind of digital information as Web content, cultural work and even personal life (friends, messages, places, etc.) [2].

The Geotopia Project (Geotopia is the French acronym for "Geolocalize to transmit, organize, share and interpret archive data") aims to use geospatial techniques for fostering a collaborative enrichment of archive documents. The project is implementing an on-line platform in order to share access, publication and annotation of archive data between archive specialists, scholars and different kind of learned people or technical experts. The project addresses questions related to geoparsing, geovisualisation and collaborative georeferencing. It is also an opportunity to experiment and observe the use of these tools in a real context. One main goal is to understand the role of georeferencing in a collaborative work [3]: metadata management and sharing, gazetteer building, collaborative work organization (by areas, topics, projects, etc.), copyrights management, collaboration between institutions, groups and individuals (amateurs) etc.

The CMS developed in the project supports digitalized archive data publishing, indexing and retrieval. Such data added value relates to local cultural heritage, and therefore, to geographic characteristics. Although well-known search engines still deliver good results of pure key-word searches, it has been observed that precision is decreasing, which in turn means that a user has to spend more time in exploring retrieved documents in order to find those that satisfy the information needed [4]. One way of improving precision is to include a geographical dimension into the search as promoted in the GeoTopia project.

During the last 20 years, open source software has undergone an impressive growth in every application domain of Computer Science (Office productivity tools, Databases, Operating Systems. . .). Among the most noticeable examples, dedicated to NLP (Natural Language Processing), are Linguastream (free of charge licence to academic users, for research purposes), TreeTagger (free of charge licence for evaluation, research and teaching purposes), Gate (licensed under LGPL) prototypes. In the mean time GIS and geo-processing frameworks have been developed (TerraLIB, JGrass, Kalipso, SAGA, SEXTANTE, OrbisGIS, PostGIS, MapServer, OpenLayers, etc.). This evolution gave a new impulse to geomatics. New projects like the GeoTopia one combine NLP tools, GIS functions and geographic resources like gazetteers to support geographic information process flows. In this paper, we are going to focus on a process flow dedicated to automatic geolocalization of spatial information embedded within texts.

Extracting different types of entities from text is usually referred to as Named Entity Recognition (NER). For over a decade, this has been an important NLP task [5]. NER has been successfully automated with near-human performance [6]. However, the work described here differs from the standard NER tasks for the following reason: the types for our geographic named entities (e.g. references to cities, streets,

rivers, mountains) are more accurate than the course-grained types that are generally considered (i.e. person, organization or location). Traditional NER systems combine lexical resources (i.e. gazetteers) with shallow processing, consisting of at least a text segmenter, a lexicon and named entities extraction rules. The extraction rules combine lexicon names with clues like capitalization and surrounding text [6]. They can be generated by hand or automatically. The former method relies on experts, while the latter uses machine learning with manually annotated training data.

As promoted by [6] for non-English languages or very specific tasks, such as the problem of handling thin-grained geographical references, we propose a process flow [7], [8] based on rules generated by hand.

In order to integrate easily such a geographical information indexing flow in the GeoTopia CMS or in existing library or document management systems, we propose to encapsulate it within the architecture of a web service: GeoStream web service. Moreover, we intend to make its behavior dynamically adaptable. For example, it might be convenient to choose the suitable gazetteer or set of gazetteers when calling the web-service. It might also be interesting to define some degree of priority for such gazetteers invocation according to the characteristics of the text to be indexed.

The paper is organized as follows. In the second section we present related work. In the third section we describe a process flow supporting automatic Geolocalization of spatial information contained within a text. Finally, in the fourth section we describe dynamic adaptation of the behavior of such a geolocalization web service

2 Related work

Automatic geolocalization of each spatial information embedded within texts relies on geographic named entity recognition. This recognition is the first step to identify the spatial information spread around.

Hereinafter related work points out spatial information is classified into at least two types of spatial information, also called ”spatial features” (SF). Simple entities correspond to named entities (e.g. ”Paris”, ”Vignemale peak”) and complex entities are derived from simple ones (e.g. ”south of Paris”, ”near Vignemale peak”). We consider simple entities to be absolute spatial features (ASF); whereas complex entities are qualified as relative spatial features (RSF) [9].

Of all the work which focuses on indexing simple entities (Cf. Table 1), only the GeoSem [10] and the PIV [9] (Virtual Itineraries in the Pyrenees) projects deal with complex entities (RSF) during the indexing and the IR phases. The Spirit prototype [11] only handles RSF during the IR phase: it provides a selection list for the user (e.g. ”north of”, ”in” etc.). In fact, all systems, except the PIV, use Minimum Bounding Rectangles (MBR) to represent SF. These produce less precise spatial representations. On the other hand, the PIV prototype uses more accurate representations (points, lines, polygons). All these systems except GeoSem, are supported by the geographical operators and functions proposed by a GIS (e.g. overlapping, intersec-

tion, etc.). A variety of spatial IR (Information Retrieval) scores based on overlaps are discussed in [12] for these prototypes. Concerning the possibilities of the querying user-interface, it is generally based on keywords. However some systems like GRID [13] propose a cartographic interface, or like STEWARD [14] propose a text field for coordinates (latitude/longitude).

Moreover, the STEWARD system [14] features synthetic views: it uses the frequency of ASF to determine the reference zone associated with each document. The GeoSem project also uses statistical approaches to process geographical queries.

Characteristics \ Prototypes	PIV_v1	STEWARD	GRID	Spirit	GéoSem
Corpus	Books	Short Text	Web Pages	Web Pages	Long Text
Document returned (document unit)	Paragraphs of a book	//	//	//	Part of a text
Complex entities (RSF)	Index+IR	/	/	IR	Index+IR
Representation	Polygons	MBR	MBR	MBR	MBR
User Interface (Kw: keywords, Carto: Cartographic Interface)	Kw and/or Carto	Kw and/or Lat/Long	Carto	Kw and/or Carto	Kw
Statistics	/	x	/	/	x
GIS	x	x	x	x	/

Table 1. Comparison of projects and prototypes used in spatial IR

The GeoStream web service we detail in sections 3 and 4 is an extension of the proposal we experimented in the PIV_v1 project.

The OGC (Open Geospatial Consortium [1]) proposes platform-independent standards defining interfaces and operations for data access and manipulation on a set of geographic features. For instance, it proposes Web Feature and Web Map Services dedicated to geographic features management within GIS. Our work is a first step for a similar contribution dedicated to geographic features management within textual digital documents.

Similarly, ViaMichelin [2], IGN GeoPortail [3], GoogleMaps [4], Geonames [5] propose web services that return an address, GPS coordinates or nearby points of interest corresponding to a toponym. Moreover, they may return an itinerary corresponding to two toponyms (departure and arrival). Few works concern services dedicated to text documents. For instance, the Geonames proposes the Wikipedia Fulltext Search web service: it returns the wikipedia entries (as xml documents) found for the searched toponym (place name). The GeoStream web service, described in this article, aims

[1] http://www.opengeospatial.org/

[2] http://www.viamichelin.fr/

[3] http://www.geoportail.fr/

[4] http://maps.google.fr/

[5] http://www.geonames.org/

to parse a textual document, to mark, analyze and geolocalyze spatial information contained in the document.

The specificities of spatial information evocation within texts, make necessary the use of different parameters and different resources like gazetteers. GeoStream uses reflection to find the right method of the right class to be performed at run time.

Reflection [15] is the process by which a computer program can observe and modify its own structure and behavior. It is a particular kind of meta-programming. Among the features provided by a language supporting reflection is the ability to convert a string matching the name of a class or function into a reference to or invocation of that class or function.

3 GeoStream: a web service supporting automatic tagging, interpretation and geolocalization of spatial information contained within a text

In this section we describe a process flow supporting automatic tagging, interpretation and geolocalization of spatial information contained within a text. Firstly, we present an example of text. Then, we briefly depict the spatial model on which GeoTopia spatial information process flow relies. Finally, we explain the main stages of the process.

3.1 Example

The following example (Figure 1) is an extract of a French text that may be translated as ”Jullian reminds of the weapon Factory of Saint-Etienne. It was settled in the Marais quarter in spring 1864.” The first stages of our information process flow mark candidate spatial and temporal features. Here (Figure 1), ”Julian”, ”Factory”, ”Saint-Etienne” and ”Marais quarter” are candidate spatial features with respective scores of 0.1; 0; 0.2 and 0.3 whereas ”spring 1864” is a candidate temporal feature. These scores (between 0 and 1) evaluate the possibility for each tagged block of text to be spatial (respectively temporal) information.

If we focus on spatial information we can mention that ”Marais quarter” is tagged with a stype = ’S’ which means that it might be a point of interest. The whole process is detailed in the two next sections.

3.2 Spatial model

The proposed semantic process for analyzing spatial features (SF) relies on an adaptative core model [7]. The model is based on a quite naive formal representation of

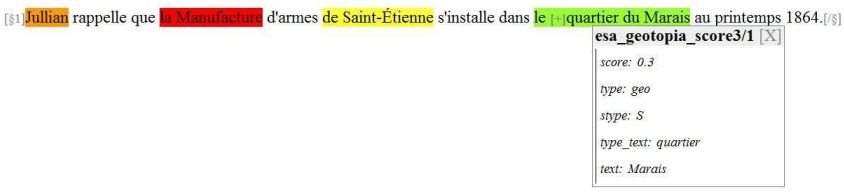

Fig. 1 An extract of a French text after making use of a lexical analysis and some spatial and temporal specific grammar rules.

spatial features in comparison with those present in the world of GIS ([16], [17], [18], GML[6] or WKT[7]). It is well appropriate to non-structured textual documents we manage in cultural heritage corpora. According to the linguistic hypothesis, the SF and the temporal features (TF) components may be recursively defined from one or several different SF and/or TF and their relationships; this idea, explained in [19], has been easily defined in a recursive way (Figure 2).

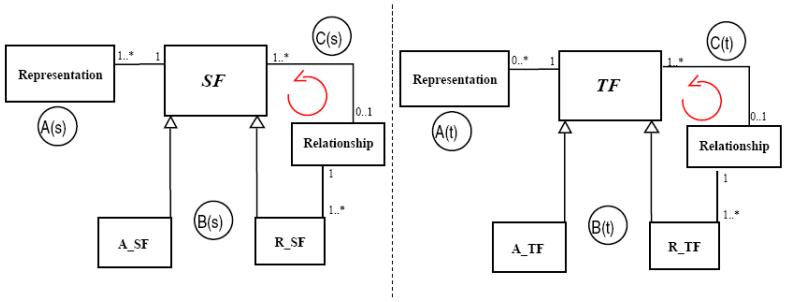

Fig. 2 Simplified diagram of spatial & temporal core models

So a SF (Figure 2) has (A(s)) at least a geometric representation. A SF could be (B(s)) an Absolute Spatial Feature (ASF), if it only consists in one named entity allowing a geolocalization. Or a SF could be a Relative Spatial Feature (RSF) if it is defined using a spatial relationship with at least one SF. (C(s)) Spatial relationships can be topological (proximity, inclusion etc.) or euclidean (distance, geometric, orientation...) [20], [21]. For instance a proximity relationship appears when we evoke a SF's spatial adjacency to another SF. This relationship is evoked in written language with terms like 'near' 'close by', etc. as "near Saint-Etienne", where the whole expression is a RSF; whereas "Saint-Etienne" is an ASF.

Every relationship is characterized by attributes in order to characterize it. For example a relationship of distance has a numerical parameter; a relationship of proximity (adjacency operator and qualifier) [12]. All the resulting SF are conform to

[6] Geography Markup Language - http://opengis.net/gml

[7] Open Geospatial Consortium - http://www.opengeospatial.org/

the spatial core model. Thus, we can manage ASF like ”in Saint-Etienne” and relative ones like ”near Saint-Etienne”, ”at about 10 km south from Saint-Etienne”, ”between Saint-Etienne and Grenoble”, etc.

Although this model has been built thanks to linguistic ideas on spatial reasoning, a similar diagram can be modeled for time reasoning [22], [23].

3.3 Spatial Information Processing

A document textual content processing sequence is usually composed of four main steps: (1) ”tokenization” divides the document into smallest blocks of text, (2) lexical and morphological analysis carries out recognition and transformation of these blocks into lexeme, (3) the syntactic analysis, based on grammars rules, allows bonds between lexeme to be found, finally, (4) the ”semantic” step carries out a more specific analysis allowing meaningful lexeme groupings to be interpreted.

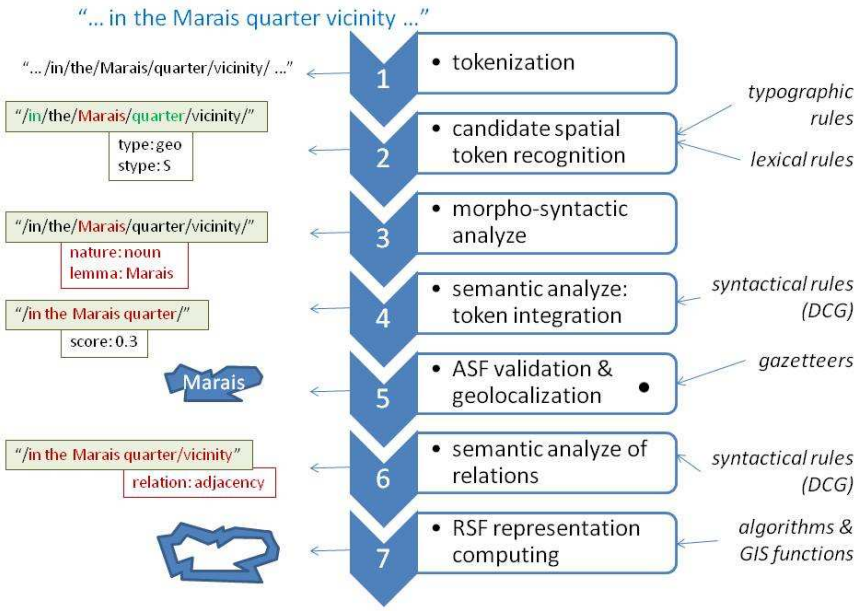

Fig. 3 Textual spatial information process flow

Our data processing sequence is a little bit different (Figure 3). After a classical preprocessing textual tokenization (1) sequence and according to [24] we adopt an active reading behavior, that is to say sought-after information is a priori known. A marker of candidate spatial token (2) locates spatial named entities using typo-

graphic and lexical rules (involving spatial features initiator lexicons). 'stype' characteristics illustrated in Figure 1 are computed during this step. Then, a morpho-syntactic analyzer (3) associates a lemma and a nature to each candidate token (i.e. "Marais", noun). A semantic analyzer (4)(6) marks candidate ASF first and candidate RSF next thanks to a Definite Clause Grammar (DCG). For instance, syntagms of composed nouns (i.e. "Marais quarter", "Emile Zola street", "Wild Chamois peak") are brought together (4) and scores are computed (i.e. figure 1).

ASF are validated and geolocalized (5) thanks to external and/or internal gazetteers (IGN French Geographic Institute resources, Geonames resources and contributive hand-craft local resources) at step five. Then multi-grained expressions containing RSF are built from pointed out ASF (6): embedded spatial relations are interpreted and corresponding geometries are computed (7).

This spatial information process flow (Figure 3) is detailed in [7] and [25]. A similar one dedicated to temporal information is detailed in [23]. Both have been experimented on samples of Pyrenean cultural heritage corpora. GeoTopia project was an opportunity to experiment this approach on new samples of Rhône-Alpes cultural heritage. For an easy integration of the process flow into the GeoTopia platform we encapsulated it within a web service. Currently, the first five stages are integrated into the GeoStream web service so that it tags and geolocalizes the ASF of texts.

Obviously, it might be interesting to indicate which set of rules, which gazetteer (set of gazetteers) is convenient or how many results of geolocalization are required at the most, etc. each time the web service is invocated. This is the reason why we propose a dynamic adaptation of the behavior of such a geolocalization web service. We describe it in the next section.

4 GeoStream: a Dynamic adaptation of the behavior of a Geolocalization Web Service

We are involved in GeoStream because our objective is to have the processes performed by the web service proposed within the GeoTopia project suitable for different but similar needs.

In addition to the text file to be parsed, GeoStream requires a second file containing a process description. This description specifies the analyses to be done and the resources to be queried. It can be seen, in a way, as a remote configuration file of the service.

4.1 GeoStream overview

GeoStream needs two input files, uses one or several resources and produces one output file. The text file to be parsed is just a text; e.g. in the following example

it's a UTF-8 text referenced by an URL. The GeoStream description document is a configuration file in xml format (Cf. 4.3). The resources can be either geographical databases or local contributive gazetteers. The result is an index in xml format (Cf. 4.3).

4.2 GeoStream Description Document (GDD)

The GeoStream Description Document states both the process to be performed and the parameters desired. The configuration of the process flow allows for permanent evolution and improvement. The basic point of this configuration is the XML-formatted string sent to GeoStream. The whole process flow is described according to this formalism where the components, their configuration, and the order in which they are invoked, are specified .

In GeoStream, the use of GDD makes it possible to choose (1) the kind of analysis (parsing for Absolute Spatial Features, Relative Spatial Features, Absolute Temporal Features or Relative Temporal Features), (2) which resources are to be consulted and, if any, in which order, (3) what the threshold given as the expected number of results is and (4) which geographic area is concerned.

With the possibilities offered by the java.lang.reflect package, the RunGeoStream (String XMLConfiguration) method, that runs GeoStream, can (1) instantiate the components specified in GDD document, (2) add them to a Stream class instance and (3) execute the flow.

Next, we examine a way to use GeoStream to parse the text showed in figure 1.

4.3 Example of running GeoStream

Here is the GeoStream description document which specifies (1) the process flow and (2) the parameters to be applied.

```
 1<?xml version="1.0" encoding="UTF-8"?>
 2 <geotopia-stream>
 3   <composant ordre="1" debug="1" class="...Composants.Input.InputURL">
 4     <param nom="Url">http://partage.clubdefrance.com/manif.txt</param>
 5   </composant>
 6   <composant ordre="2" debug="1"
 7     class="...Composants.Linguastream.Linguastream">
 8     <param nom="Ls_stream_base_dir">.../geotopia_WS_package/lib/chaine_ls/
       </param>
 9     <param nom="Ls_stream_file">esa.ls</param>
10   </composant>
11   <composant ordre="3" debug="1"
12     class="...Composants.Georeferenceur.Georeferenceur">
13       <param nom="XpathES">//es</param>
14       <param nom="XpathESText">./text/text()</param>
15       <ressources>
16       <ressource class="fr ... Composants.Georeferenceur.PostgresGazetteer"
                        ordre="1">
17       <!-- IGN resource IGN limited to Pyrenees-Atlantiques-->
18         <param nom="Source">Communes IGN</param>
```

```
19          <param nom="Host">postgis.clubdefrance.fr</param>
            ...
23          <param nom="Query">select nom_com as NOM, 'P' as TYPE,
              astext(the_geom) as WKT
              from communes_ign where (nom_com ilike '_-_NOM_-_'
              and  intersect(the_geom, GeomFromText('POLYGON(...))'))</param>
24        </ressource>
25        <ressource class="fr ... Composants.Georeferenceur.Geonames" ordre="1">
28        <!--ressource Geonames.org limited to France, seuil=0-->
29          <param nom="Parametre">country=FR</param>
30          <param nom="Seuil">0</param>
31        </ressource>
32      </ressources>
33    </composant>
34    <composant ordre="4" debug="1"
35      class="fr.unipau.liuppa.geotopia.Composants.XSLTransfo.XSLTransfo">
36      <param nom="XslFilePath">/opt/geotopia_WS_package/lib/xsl/geotopia.xsl</param>
37    </composant>
38 </geotopia-stream>
```

In this example, we notice two resources. The first resource, as can be seen in line 16, is a contributive homemade gazetteer (supported by Postgis). The second one refers to Geonames.

This document specifies orders (lines 3, 6 , 11, 16), thresholds (line 30) and geographic area (lines 23, 29). The components are considered in the order fixed by numbers (e.g. ordre=”2”). The number of results needed to skip other unexamined resources is given by the threshold (e.g. Seuil$>$0). The geographic area can be delimited either by a parameter (e.g. country=France) or by restriction in the sql query (where . . .).

With the description document, given above, we parse the example (Cf. fig 1) and we obtain the following file:

```
<?xml version="1.0" encoding="UTF-8"?>
<geotopia-doc xsi:noNamespaceSchemaLocation=
   "http://geotopia.univ-pau.fr:8180/schema/geotopia-doc.xsd"
   doc_source="_-_DOC_BASE_-_" doc_original="_-_DOC_ORIGINAL_-_"
   date_interpretation="2009-03-30"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<doc>
   <paragraphe id="1">
      <token id="1"><text>Jullian</text></token>
      <token id="2"><text>rappelle</text></token>
      ...
      <token id="6"><text>Manufacture</text></token>
      <token id="7"><text>d'</text></token>
      <token id="8"><text>armes</text></token>
      <token id="9"><text>de</text></token>
      <token id="10"><text>Saint-Etienne</text></token>
      ...
      <token id="15"><text>quartier</text></token>
      <token id="16"><text>du</text></token>
      <token id="17"><text>Marais</text></token>
      <token id="18"><text>au</text></token>
      <token id="19"><text>printemps</text></token>
      <token id="20"><text>1864</text></token>
      <token id="21"><text>.</text></token>
   </paragraphe>
</doc>
<sem>
   ...
   <es id="2" id_paragraphe="1" id_token_debut="9" id_token_fin="10">
      <text>Saint-Etienne</text><type>null</type><poids>0.2</poids>
      <georeferencement source="Communes IGN" query=
```

```
            "select+nom_com+as+NOM~+'P'+as+TYPE~+astext(the_geom)+as+WKT
             +from+communes_ign+where+(nom_com+ilike+'Saint-Etienne')"
             geotype="P" geoname="SAINT-ETIENNE">
            <geodata>MULTIPOLYGON(((766195.904391206 2050603.2318322,766217.901342646
                  2050519.02475247,766313.336033008 2050396.78127754,766475.79268479
                  ...
                  2049504.52993475,749543.181043979 2049903.56838197,749497.354061812
                  2049973.68366468,749263.636452761 2050331.36326049)))
           </geodata>
        </georeferencement>
    </es>
    <es id="3" id_paragraphe="1" id_token_debut="14" id_token_fin="17">
       <text>Marais</text><type>S</type><poids>0.3</poids>
       <georeferencement source="geonames" query=
          "http://ws.geonames.org/search?style=SHORT&name=Marais&country=FR"
          geotype="P"   geoname="La Chapelle-des-Marais">
       <geodata>POINT(47.45 -2.25)</geodata>
       </georeferencement>
       ...
       <georeferencement source="geonames" query=
          "http://ws.geonames.org/search?style=SHORT&name=Marais&country=FR"
          geotype="A" geoname="Ponts-et-Marais">
          <geodata>POINT(50.05 1.45)</geodata>
       </georeferencement>
    </es>
</sem>
</geotopia-doc>
```

This file has two parts: the first one tagged by <doc> and </doc> reports the tokenization, the second one tagged by <sem> and </sem> lists the recognized spatial features with their localizations.

It's noticeable that Saint-Etienne has been found as an absolute spatial feature and is returned with its geometry.

The research for "quartier du Marais" has failed so it has been extended to "Marais" which has given many results.

4.4 Use of java.lang.reflect

As said in the Java documentation, the java.lang.reflect package provides classes and interfaces for obtaining reflective information about classes and objects. The following program uses the class method defined in this package in the same way as the RunGeoStream method.

```
 1 import java.lang.reflect.*;
 2 public class  FooTest{
 3   public static void main (String a [ ])
 4     throws ClassNotFoundException,
 5            NoSuchMethodException,
 6            IllegalAccessException,
 7            InstantiationException,
 8            InvocationTargetException
 9            {
10     String className = a[0];
11     String methodName = a[1];
12     Class parameterType [ ] = new Class [a.length - 2];
13     String argument [ ]      = new String [a.length - 2];
14     for (int i=2; i<a.length; i++) {
15       argument[i-2] = a[i];
```

```
16       parameterType[i-2] = argument[i-2].getClass();
17     }
18
19     // using java.lang.reflect
20     Class cl = Class.forName (className);
21     Method method = cl.getMethod (methodName, parameterType);
22     method.invoke (cl.newInstance (), argument);
23
24     System.out.println("End of FootTest");
25   }
26 }
```

The three lines, which perform the dynamic feature of GeoStream, have to be explained :

- in line 20, the 'forName' method of the predefined 'Class' class returns the class of the argument given by 'className' into the 'cl' variable.
- in line 21, the 'getMethod' method is used to find the method that suits the 'methodName' and 'parameterType' arguments.
- in line 22 the previous appropriate method is sent to an instance of the cl class with the relevant arguments.

Two examples of use are shown below:

```
> java FooTest Foo react1 a_string
Method react1 is running, the parameter is "a_string"
End of FootTest

> java FooTest Foo react2 string1 string2
Method react2 is running, the parameters are "string1" and "string2"
End of FootTest
```

for the given 'Foo' class:

```
public class Foo {
  ...
  public void react1(String S) {
    System.out.println
      ("Method react1 is running, the parameter is \""+S+"\"");
  }
  public void react2(String S1, String S2) {
    System.out.println
      ("Method react2 is running, the parameters are \""+S1+"\" and \""+S2+"\"");
  }
}
```

5 Conclusion

The problem of most current CMS integrating the geographic dimension for document indexation and retrieval is that they usually need manual annotations to associate one or more spatial footprints to a document. In this paper we describe an approach dedicated to the automatic indexation and geolocalization of spatial information embedded within texts . We propose and experiment the corresponding GeoStream web service within the GeoTopia CMS. This web service tags, interprets and geolocalizes spatial named entities we call ASF.

We complete this web service behavior with dynamic adaptation possibilities. In such a geographic context, during the indexing stage, it is very interesting to dynamically associate a specific analyzing process flow and specific required resources to each new document stored in the CMS. We use the java.lang.reflect package to experiment and validate reflection through different GDD files specifications.

We extend this spatial information management process flow to the management of complex features (RSF) [11]. Future improvement of the GeoStream web service would be to propose and experiment a new version incorporating the management of RSF, of different text formats (UTF-8, ISO-8859-15, etc.).

The final goal is the dynamic choice and running of a subset of steps for the information process flow. As the current choice is limited to ASF, it would be interesting to choose whether to parse ASF, RSF, ATF or RTF (Temporal Features) dynamically in a next GeoStream version.

Acknowledgements GeoTopia project is funded by the CNRS-TGE-Adonis (NO SUBOS-20.DR7). It is led in partnership with the EVS (umr5600.univ-lyon3.fr/) and LIUPPA (http://liuppa.univ-pau.fr) laboratories and the Makina Corpus (http://www.makina-corpus.com/) company.

Glossary

ASF	Absolute spatial Feature
ATF	Absolute temporal feature
CMS	Content management system
DCG	Definite clause grammar
GDD	Geostream description document
GIS	Geographic information system
IR	Information retrieval
MBR	Minimum bounding rectangles
NER	Named entity recognition
NLP	Natural language processing
PIV	French acronym for Virtual Itineraries in the Pyrenees Mountains
RSF	Relative spatial feature
RTF	Relative temporal feature
SF	Spatial feature
TF	Temporal feature

References

1. Hill L. (2006) Georeferencing, The Geographic Associations of Information. Boston, The MIT Press. 272
2. Joliveau T. (2009) Connecting Real and Imaginary Places through Geospatial Technologies: Examples from Set-Jetting and Art-Oriented Tourism, Cartographic Journal Vol. 46 (Issue 1, Cinematic cartography,): (in press)
3. Tuffery C, Fernandes P, Le Ny P-A (2008) Utilisation d'un site web intégré de webmapping et de gestion de contenus pour la publication de ressources documentaires géoréférencées, Colloque international Le SIG WebMapping dans les sciences archéologiques et historiques, Paris

4. Kanhabua N, Nørvåg K (2008) Improving Temporal Language Models for Determining Time of Non-timestamped Documents, B. Christensen-Dalsgaard, D. Castelli, B.A. Jurik, J. Lippincott, ECDL. Vol. 5173 of Lecture Notes in Computer Science, 358–370
5. Chinchor N (editor) (1998) Proceedings of the 7th Message Understanding Conference
6. Martins B, Manguinhas H, Borbinha J (2008) Extracting and Exploring the Geo-Temporal Semantics of Textual Resources, IEEE International Conference on Semantic Computing, IEEE DOI 10.1109/ICSC.2008.86
7. Gaio M, Sallaberry C, Etcheverry P, Marquesuzaa C, Lesbegueries J (2008) A global process to access documents' contents from a geographical point of view, Journal of Visual Languages And Computing. Vol. 19., Orlando, FL, USA, Academic Press, Inc. 3-23
8. Sallaberry C, Baziz M, Lesbegueries J, Gaio M (2007) Towards an ie and ir system dealing with spatial information in digital libraries - evaluation case study, J. Cardoso, J. Cordeiro, J. Filipe, ICEIS 5:190-197
9. Lesbegueries J, Gaio M, Loustau P, Sallaberry C (2006) Geographical information access for nonstructured data. In Haddad, H., ed.: SAC, ACM 83-89
10. Bilhaut F, Dumoncel F, Enjalbert P, Hernandez N (2007) Indexation sémantique et recherche d'information interactive. In: CORIA 2007. 65-76
11. Vaid S, Jones C B , Joho H, Sanderson M (2005) Spatio-textual indexing for geographical search on the web. In Medeiros, C.B., Egenhofer, M.J., Bertino, E., SSTD. Vol. 3633 of Lecture Notes in Computer Science., Springer 218-235
12. Sallaberry C, Gaio M, Palacio D, Lesbegueries J (2008) Extending gis functions for gir needs. In: ACM 17th CIKM, (GIR workshop), Napa Valley, CA. 1-8
13. Valcartier (2006) Grid - geospatial retrieval of indexed document. Technical report, R&D pour la défense Canada
14. Lieberman M D, Samet H, Sankaranarayanan J, Sperling J (2007) Steward : architecture of a spatio-textual search engine. In: GIS '07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, New York, NY, USA, ACM 1-8
15. Sobel J M, Friedman D P (1996) An introduction to reflection-oriented programming, Proceedings, Reflection 96, San Francisco, 107–126.
16. Clementini E, Sharma J, Egenhofer M J (1994) Modeling topological spatial relations: Strategies for query processing, Computers and Graphics 18 (6):815-822
17. Egenhofer M J (2002) Toward the semantic geospatial web, GIS'02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems, ACM Press, 1-4
18. Hill L L (2000) Core elements of digital gazetteers: Placenames, categories, and footprints ECDL'00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. 280–290
19. Vandeloise C (1986) L'espace en français, Travaux Linguistiques
20. Cohn A G (1997) Qualitative spatial representation and reasoning techniques, KI '97: Proceedings of the 21st Annual German Conference on Artificial Intelligence, 1–30
21. Cohn A G, Hazarika S M (2001) Qualitative spatial representation and reasoning: An overview Fundamenta Informaticae, 46(1-2):1–29
22. Bessière C, Euzenat J, Jeansoulin R, Ligozat G, Schwer S (1997) Raisonnement spatial et temporel Actes 6e journées nationales du PRC-GDR intelligence artificielle,
23. Parc-Lacayrelle A L, Gaio M, Sallaberry C (2007) La composante temps dans l'information géographique textuelle, Revue Document Numérique 10(2):129-148
24. Baccino T, Pynte J (1994) Spatial coding and discourse models during text reading, Language and Cognitive Processes, 9:143-155
25. Loustau P (2008) Interprétation automatique d'itinéraires dans des récits de voyages. D'une information géographique du syntagme à une information géographique du discours Thèse de doctorat, soutenue à l'Université de Pau et des Pays de l'Adour